# A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets

Rémi Bardenet, Odalric-Ambrym Maillard

# A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets

**Rémi Bardenet** [1,2], **Odalric-Ambrym Maillard** [1,3]

[1] *Both authors contributed equally to this work.*

[2] *CNRS & CRIStAL, Univ. Lille, France*
*e-mail:* remi.bardenet@gmail.com

[3] *Team TAO, Inria Saclay - Île de France & LRI, France*
*e-mail:* odalricambrym.maillard@inria.fr

**Abstract:** New Markov chain Monte Carlo (MCMC) methods have been proposed to tackle inference with tall datasets, i.e., when the number $n$ of data items is intractably large. A large class of these new MCMC methods is based on randomly subsampling the dataset at each MCMC iteration. We investigate whether random projections can replace this random subsampling for linear regression of big streaming data. In the latter setting, random projections have indeed become standard for non-Bayesian treatments. We isolate two issues for MCMC to apply to streaming regression: 1) a resampling issue; MCMC should access the same random projections across iterations to avoid keeping the whole dataset in memory and 2) a budget issue; making individual MCMC acceptance decisions should require $o(n)$ random projections. While the resampling issue can be satisfyingly tackled, current techniques in random projections and MCMC for tall data do not solve the budget issue, and may well end up showing it is not possible.

## 1. Introduction

Markov chain Monte Carlo algorithms (MCMC; Robert & Casella, 2004) are loops that sweep over the whole dataset at each iteration, thus making the application of such methods to large datasets delicate. The flagship of MCMC is the Metropolis-Hastings algorithm (MH), which is very popular for Bayesian inference in complex models (Brooks *et al.* , 2011, Part II). With the recent *big data* and *fast data* revolutions, a lot of effort has been recently spent on proposing variants of MH that scale to *tall* datasets, when the bottleneck is the number $n$ of individual items while the dimension of the data remains tractable for MCMC, see (Bardenet *et al.* , 2015) for a review.

On the other hand, random matrix theory has generated a paradigm shift in the machine learning community over the last decade, providing well-grounded dimension reduction tools at a cheap computational price. Random projection techniques have been successfully applied to large dimension problems in compressed sensing (Donoho, 2006; Candes & Tao, 2006; Baraniuk, 2007), regression (Maillard & Munos, 2012; Zhang *et al.* , 2013), clustering (Sakai & Imiya, 2009), classification (Duarte *et al.* , 2007) or sparse representations (Hamilton *et al.* , 2013) to name a few. This success is based on concentration inequalities (Dasgupta & Gupta, 2003), together with low numerical cost implementations (Achlioptas, 2003). Importantly, algorithms that rely on four-wise independent Johnson-Lindenstrauss matrices (Alon *et al.* , 1996, 1992, 2007; Wang *et al.* , 2007) achieve high numerical performance for streaming data, even under constraints such as turnstile updates, when

data arrives in the form of statements like "add quantity $q$ to the $(i, j)$th element of the data matrix".

Streaming MCMC would be MCMC-like algorithms that require only one pass over the data, $o(n)$ storage, and still perform sampling correctly with large probability. To achieve this, we ask to which extent random projection techniques can be combined with MH.

**Previous work.** Variants of MH for tall datasets can be classified in two groups. *Divide-and-conquer* approaches partition the dataset into manageable parts, run MCMC on each part, and combine the results of these multiple runs (Huang & Gelman, 2005; Scott *et al.* , 2013; Neiswanger *et al.* , 2014; Minsker *et al.* , 2014; Srivastava *et al.* , 2014). *Subsampling* approaches rely on randomly subsampling the dataset at each MCMC iteration (Welling & Teh, 2011; Bardenet *et al.* , 2014; Korattikara *et al.* , 2014; MacLaurin & Adams, 2014; Quiroz *et al.* , 2014, 2015). In this paper, we focus on the subsampling approach of Bardenet *et al.* (2014), henceforth "confidence MH", for three reasons. First, it relies on concentration inequalities, which is precisely the kind of guarantee random projections provide. Second, confidence MH inherits the uniform ergodicity of the underlying intractable MH. In other words, one iteration of confidence MH is as efficient as one iteration of an MH with the same user-defined parameters. We can thus make sense of the computational cost of one iteration of confidence MH, since roughly the same total number of iterations as MH is necessary for a given accuracy. Third, the price of these theoretical guarantees is a high computational cost precisely because of subsampling and conservative concentration inequalities, so that random projections can potentially benefit this algorithm the most.

Regarding the *budget issue*, that is, the number of datapoints needed at each iteration of the algorithm in order to maintain performance guarantees, Bardenet *et al.* (2014) explained why confidence MH requires $\Theta(n)$ datapoints per iteration to guarantee the inheritance of the convergence speed of the underlying, intractable MH. This cost was then lowered to $o(n)$ in Bardenet *et al.* (2015), using control variates such as Taylor expansions around the maximum likelihood estimator (MLE), assuming these are available. At this stage, we note that if data is streaming and arrives in a turnstile manner (Muthukrishnan, 2005; Clarkson & Woodruff, 2009), even the MLE might not be easy to compute, even for simple models. Additionally, confidence MH suffers from a *resampling issue*: drawing a different random subsample of the dataset at each MCMC iteration requires the ability to store and repeatedly query the full dataset.

Finally, Geppert *et al.* (2015) have applied random projections to Bayesian regression of tall data, recommending the use of the posterior of the projected dataset, built by a single multiplication by a random matrix. Good experimental results have been shown; we argue in Section 4.3, however, that the accuracy guarantee of Geppert *et al.* (2015) is too weak for some important Bayesian inference tasks.

**Contribution.** We combine existing results from distinct communities, namely the MCMC and random projections literature, to investigate whether random projections can advantageously replace uniform subsampling in the confidence MH of Bardenet *et al.* (2014). For that purpose and to gain as much intuition on the problem as possible, we focus on the illustrative case of Bayesian linear regression. This is arguably one of the simplest tasks that MCMC can be applied to, which still captures the difficulty for MCMC to handle tall datasets. Furthermore, it is amenable to powerful random projection techniques, also known as *sketching* (Clarkson & Woodruff, 2009).

Our answer to the title question is yes and no: while random projections successfully tackle the resampling issue, the budget issue seems to be hard to overcome. Overall, we isolate the strengths and weaknesses of random projections as a tool for subsampling-based MH.

In Section 2, we introduce subsampling-based Metropolis-Hastings for regressing tall data, and

detail the confidence MH algorithm of Bardenet *et al.* (2014). On the positive side, in Section 3, we use a uniform concentration result of Sarlós (2006) to build a candidate projection-based confidence MH that can address the resampling issue. In Section 4, we take a close look at the confidence bounds provided by random projection techniques, and argue that solving the budget issue is not possible with typical random projection concentration inequalities. In particular, we argue that a naive application of the control variate technique that solves the budget issue for confidence MH in (Bardenet *et al.* , 2015) fails here.

## 2. Metropolis-Hastings for regressing tall data

Consider a regression dataset of $n$ points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$ arranged in a $n \times d$ matrix $X$, along with the corresponding regressed variable $Y \in \mathbb{R}^n$. Given a prior $p(\theta)$ on the regression coefficients $\theta \in \mathbb{R}^d$, Bayesian linear regression assumes a generative model $Y|X, \theta \sim \mathcal{N}(X\theta, \sigma^2 I)$. For ease of derivation, we will consider here the noise $\sigma$ to be known and equal to $1/\sqrt{2}$, so that the posterior reads

$$\pi(\theta) \propto \gamma(\theta) = \exp -\|X\theta - Y\|^2 p(\theta). \tag{1}$$

Extending our algorithm to an unknown $\sigma$ will be straightforward. For general priors, the posterior (1) is intractable and we rely on Bayesian computational tools to approximate it.

A standard approach to sample approximately from $\pi$ is the Metropolis-Hastings algorithm (MH; Robert & Casella, 2004, Chapter 7.3). MH consists in building an ergodic Markov chain of invariant distribution $\pi$. Given a proposal $q(\theta'|\theta)$, MH starts its chain at a user-defined $\theta_0$, then at iteration $k+1$ it proposes a candidate state $\theta' \sim q(\cdot|\theta_k)$ and sets $\theta_{k+1}$ to $\theta'$ with probability $\min[1, \alpha(\theta_k, \theta')]$, where

$$\alpha(\theta, \theta') \quad = \quad \frac{\gamma(\theta')}{\gamma(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}, \tag{2}$$

while $\theta_{k+1}$ is otherwise set to $\theta_k$. In practice, this last step is implemented by drawing a uniform $u \sim \mathcal{U}_{[0,1]}$ and comparing $\alpha(\theta_k, \theta')$ to $u$. Each iteration of MH thus requires to evaluate the unnormalized posterior $\gamma$ at a new point, which requires the whole dataset $(X, Y)$ by definition (1). When the data is *tall*, i.e. the number $n$ of data points is large, MH is thus often ruled as too costly. However, recent advances have been made in adapting MH to the tall data setting, see (Bardenet *et al.* , 2015) for a review.

### 2.1. Confidence Metropolis-Hastings

We focus here on a concentration-based MH proposed by Bardenet *et al.* (2014) and further improved by Bardenet *et al.* (2015). The algorithm samples a Markov chain similarly to MH, and we define it here by going through one of its iterations. Assume the current state of the chain is $\theta$ and a proposal $\theta'$ has been drawn from $q(\cdot|\theta)$, as well as a uniform $u \sim \mathcal{U}_{[0,1]}$. MH would require to compute $\alpha(\theta, \theta')$ in (2) and compare it to $u$, but we avoid this as follows. Upon noting that the average log likelihood is

$$\ell(\theta) = -\frac{1}{n}\|X\theta - Y\|^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^T\theta - y_i)^2,$$

one can build an estimate of $\ell(\theta)$ by uniformly subsampling $t < n$ rows $x_1^*, \dots, x_t^*$ of $X$ and the corresponding rows of $Y$, namely

$$\widehat{\ell}_t(\theta) = \frac{1}{t} \sum_{i=1}^{t} \left( x_i^{*T} \theta - y_i^* \right)^2 . \tag{3}$$

One can then estimate $\alpha$ in (2) through

$$\widehat{\log \alpha}_t(\theta, \theta') = n \left[ \widehat{\ell}_t(\theta') - \widehat{\ell}_t(\theta) \right] + \log \frac{q(\theta|\theta')p(\theta')}{q(\theta'|\theta)p(\theta)}.$$

Given a confidence level $\delta > 0$, concentration inequalities such as Hoeffding's (Boucheron *et al.* , 2013) yield confidence bounds $c_t(\delta)$ such that

$$\mathbb{P}\left( n^{-1} |\widehat{\log \alpha}_t(\theta, \theta') - \log \alpha(\theta, \theta')| \leq c_t(\delta) \right) \geq 1 - \delta, \tag{4}$$

where the probability $\mathbb{P}$ is over the subsampled dataset. Let $(t_i)_{i \geq 0}$ be a divergent nondecreasing sequence and $(\delta_i)$ be such that $\sum_i \delta_i \leq \delta$. In confidence MH, one increases $i$ and draws new data items until

$$n^{-1} |\widehat{\log \alpha}_{t_i}(\theta, \theta') - \log u| > c_{t_i}(\delta_i). \tag{5}$$

Denoting by $T$ the smallest $t_i$ such that (5) holds, Bardenet *et al.* (2014) show that accepting the point $\theta'$ if and only if $\widehat{\log \alpha}_T(\theta, \theta') > \log u$ amounts to the same acceptance decision as MH with probability $1 - \delta$. If the ideal, intractable MH is uniformly ergodic, then confidence MH has a limiting distribution that is within $\mathcal{O}(\delta)$ of $\pi$, and it is uniformly ergodic with close constants to the ideal algorithm (Bardenet *et al.* , 2014, Proposition 3.2), thus making one iteration as efficient as one iteration of the underlying intractable MH.

Outside requiring confidence bounds $c_t(\delta)$, confidence MH has two main disadvantages: first, Bardenet *et al.* (2014) show that $c_t(\delta)$ has to be of order $n^{-1}$ on average, which leads to the required number of subsamples $T$ to be of order $n$ at a given iteration, thus obtaining no more than a constant gain when compared to MH. We refer here to this pitfall as the *budget* issue. Bardenet *et al.* (2015) introduce control variates using Taylor expansions of the log likelihood, and use concentration inequalities to bound only the Taylor remainder of the log likelihood ratio. They achieve as low as $\mathcal{O}(1)$ data points per iteration in favourable cases. Second, confidence MH requires (4) to hold at each MH iteration independently, which means a new random subsample of the dataset has to be drawn at each MH iteration. This is also a drawback, since for numerous big data applications, storing the data and repeatedly subsampling it is not tractable. We call this pitfall the *resampling* issue. Note that this is not specific to using uniform subsampling to build acceptance ratio estimates: all MCMC algorithms with randomized acceptance steps need to guarantee some independence across iterations of the coin tosses generating the acceptance ratio estimates. In the remaining section, we investigate whether random projections can advantageously replace uniform subsampling in providing estimates of the log likelihood ratio with confidence guarantees such as (4).

## 3. A confidence Metropolis-Hastings with random projections

The log likelihood ratio corresponding to (1) is

$$\begin{aligned}\rho(\theta,\theta') &\triangleq n\left[\ell(\theta')-\ell(\theta)\right]\\ &= \|X\theta-Y\|^2-\|X\theta'-Y\|^2 \qquad (6)\\ &= \langle X(\theta-\theta'), X(\theta+\theta')-2Y\rangle. \qquad (7)\end{aligned}$$

As such, it is preserved by random projections, which preserve Euclidean norms by the celebrated Johnson-Lindenstrauss lemma (Johnson & Lindenstrauss, 1984; Boucheron *et al.* , 2013). In practice, random projections amount here to choosing a projection dimension $k \leq d$, and drawing a $k \times n$ random matrix $A$ such that replacing $X$ and $Y$ by $AX$ and $AY$ in (6) or (7) results in a controlled approximation of the log likelihood ratio. In this section, we formalize this argument and show how this yields a variant of the confidence MH in Section 2.1.

### *3.1. A uniform Johnson-Lindenstrauss lemma*

We first state the original Johnson-Lindenstrauss (JL) lemma.

*Lemma* 1 (Johnson & Lindenstrauss, 1984). Let $x \in \mathbb{R}^n$ and $\epsilon \in (0,1)$. Let $A$ be an $k \times n$ matrix of i.i.d. Gaussian $\mathcal{N}(0,1/k)$ entries. It holds

$$\mathbb{P}\left(\left|\|Ax\|^2-\|x\|^2\right| \geq \epsilon\|x\|^2\right) \leq 2e^{-k(\epsilon^2/4-\epsilon^3/6)}.$$

*Remark.* The Gaussian $\mathcal{N}(0,1/k)$ distribution in Lemma 1 can be replaced by a Rademacher $\pm 1/\sqrt{k}$, or a distribution with values $\pm\sqrt{3/k}$ with probability 1/6 and 0 with probability 1/3, see (Achlioptas, 2003). Recent works (Alon *et al.* , 2007; Ailon & Liberty, 2009; Gilbert *et al.* , 2007) have also considered more computationally efficient distributions using Rademacher matrices enjoying a 4-wise independence property, leading to faster solutions with low memory requirement. In the sequel, we refer to matrices that satisfy a JL lemma as *Johnson-Lindenstrauss matrices*.

Lemma 1 gives a confidence bound as in (4), which is only valid for one choice of $\theta, \theta'$. This is precisely what causes the resampling issue in confidence MH. To address this, we need to guarantee that the same projections $(AX, AY)$ can be reused in further MH iterations. A naive union bound argument over the whole Markov chain $(\theta_k)_{1\leq k\leq T}$ would not be valid, as the accepted states depend on $A$. Fortunately, Sarlós (2006) showed that the JL lemma can be made uniform over the span of the columns of $Z = [X, Y]$, the $n \times (d+1)$ concatenation of $X$ and $Y$. We restate the result of Sarlós (2006) here and make its confidence bound explicit.

**Proposition 3.1** (Sarlós, 2006). *Let $r = rk([X,Y])$ and $A$ be a $k \times n$ Johnson-Lindenstrauss matrix. Let $U$ be $n \times r$ with orthonormal columns. Then it holds*

$$\begin{aligned}&\mathbb{P}\left(\forall v, w \in \mathbb{R}^r, |\langle AUv, AUw\rangle - \langle v, w\rangle| \leq \epsilon\|v\|\|w\|\right)\\ &\qquad \geq 1 - 4\left[r^2 + (34r)^{2r-2}\right] e^{-k\left(\frac{\epsilon^2}{5}-\frac{\epsilon^3}{8}\right)}. \qquad (8)\end{aligned}$$

*Remark.* In this proposition, note that the high probability event holds uniformly over all $v, w \in \mathbb{R}^r$, and not only pointwise.

### 3.2. The resampling issue

To see why Proposition 3.1 yields a uniform control of the log likelihood ratio (7) on $\Theta$, let us write the SVD of $Z = [X, Y]$ as $Z = U\Sigma V^T$. Upon rewriting (7) as

$$\left\langle U\Sigma V^T \begin{pmatrix} \theta - \theta' \\ 0 \end{pmatrix}, U\Sigma V^T \begin{pmatrix} \theta + \theta' \\ -2 \end{pmatrix} \right\rangle,$$

Proposition 3.1 yields a confidence bound on the difference between (7) and

$$\langle AX(\theta - \theta'), AX(\theta + \theta') - 2AY \rangle.$$

Thus, the log likelihood ratio (7) satisfies

$$\forall \theta, \theta', \rho(\theta, \theta') \in [b_-(\theta, \theta'), b_+(\theta, \theta')] \tag{9}$$

with probability given by (8), where

$$\begin{aligned} b_\pm(\theta, \theta') &= \langle AX(\theta - \theta'), AX(\theta + \theta') - 2AY \rangle \\ &+ \frac{\epsilon}{(1 \mp \epsilon)^2} \|AX(\theta - \theta')\| \|AX(\theta + \theta') - 2AY\|. \end{aligned} \tag{10}$$

Since (9) is valid for all $\theta, \theta' \in \mathbb{R}^d$, using it in (4) allows the user to reuse the same projections $AX, AY$ from one MH iteration to the next. We give a one-pass MCMC-like algorithm in Figure 1 that relies on the uniform bound of Sarlós (2006). This algorithm potentially solves the resampling issue introduced in Section 1, as we do not need to query the dataset at each new iteration for the concentration inequality to hold. Still we are guaranteed RPMH makes the same acceptance decisions as MH with large probability, which is key for convergence properties to be derived as in (Bardenet *et al.* , 2014). Note that in confidence MH in Section 2.1, the number of subsamples is increased until the acceptance decision is made. This corresponds here to increasing the number $k$ of random projections and decreasing $\epsilon$ accordingly to control the probability (8) until the confidence interval (9) is small enough that (5) holds. We replaced this increase of $k$ by the FAIL flag in RPMH, which is raised as soon as $k$ does not allow to make a decision. We will see this simple algorithm already illustrates the weakness of the approach.

For the moment, we turn to the budget issue and investigate whether $k = o(n)$ random projections are enough for the algorithm RPMH in Figure 1 to return FAIL only with small probability over the coin tosses of the algorithm.

## 4. The budget issue

In this section, we denote by $\theta^\star$ the MLE, $\theta^\star = \arg\min_\theta \|X\theta - Y\|$, corresponding to noise $\eta^\star = X\theta^\star - Y$. Given a JL matrix $A$, we also denote by $\tilde{\theta}$ the MLE of the projected regression problem $\tilde{\theta} = \arg\min_\theta \|AX\theta - AY\|$, with corresponding noise $\tilde{\eta} = X\tilde{\theta} - Y$.

### 4.1. A close look at confidence bounds

In this section, we investigate how the size of the confidence interval (9) scales with $\epsilon$ and $n$ for different applications of Proposition 3.1, and examine the impact on the algorithm in Figure 1.

NaiveOnePassRPMH($p(\theta)$, $q(\theta'|\theta)$, $\theta_0, X, Y, A, T$)

1 Compute $AX, AY$ ▷ *can be done while acquiring data.*
2 **for** $k = 1, \dots, T$,
3 $\theta \leftarrow \theta_{k-1}$
4 $\theta' \sim q(.|\theta)$, $u \sim \mathcal{U}_{(0,1)}$,
5 $\psi \leftarrow \log\left[u \frac{p(\theta)q(\theta'|\theta)}{p(\theta')q(\theta|\theta')}\right]$
6 $\widehat{\log \alpha} \leftarrow \|AX\theta - AY\|^2 - \|AX\theta' - AY\|^2$
7 Compute $b_\pm(\theta, \theta')$ as in (10) or (13). ▷ *Compute bounds*
8 **if** $\psi(u, \theta, \theta') \in [b_-(\theta, \theta'), b_+(\theta, \theta')]$ ▷ *We cannot make a decision*
9 **return** Fail.
10 **if** $\widehat{\log \alpha} > \psi$
11 $\theta_t \leftarrow \theta'$ ▷ *Accept*
12 **else** $\theta_t \leftarrow \theta$ ▷ *Reject*
13 **return** $(\theta_k)_{k \geq 1}$

Figure 1. *The pseudocode of our proposed MH with random projections.*

**The naive approach.** As given in (10), the confidence interval (9) is centered at

$$\begin{aligned} &\langle AX(\theta - \theta'), AX(\theta + \theta') - 2AY \rangle, \\ &\leq \|AX(\theta - \theta')\| \|AX(\theta + \theta') - 2AY\| \end{aligned} \tag{11}$$

and we claim that at equilibrium of the chain, that is, when $\theta, \theta'$ are drawn from $\pi(\theta)q(\theta'|\theta)$, this middle will be within $\mathcal{O}_p(\sqrt{n})$ of 0 in most usual settings, further being negative in a significant proportion of iterations, since the proposal in MCMC for $d > 3$ is usually tuned to reach a constant acceptance around 25%. Indeed, the target (1) is expected to be close to a Gaussian

$$\pi \approx \mathcal{N}(\theta^\star, (X^T X)^{-1}), \tag{12}$$

with equality for a flat prior. Following Roberts & Rosenthal (2001), a natural choice for the MCMC proposal distribution $q(\theta'|\theta)$ is then a Gaussian centered at $\theta$ with inverse covariance proportional to $(AX)^T(AX)$.

This leads to $\|AX(\theta - \theta')\| = \mathcal{O}_p(1)$, for a given $\theta$. Now, at equilibrium of the chain, it holds

$$\begin{aligned} \|AX\theta' - AY\|^2 &\leq \mathcal{O}_p(1) + \|AX\theta - AY\|^2 \\ &\leq \mathcal{O}_p(1) + (1 + \epsilon)\left[\|\eta^\star\|^2 + \|X(\theta - \theta^\star)\|^2\right], \end{aligned}$$

so that the last term in the right-hand side of (11) is $\mathcal{O}_p(\sqrt{n})$. Note, moreover, that for "reasonable" datasets, say $X, Y$ are such that the noise $\eta^\star$ is Gaussian, the right-hand side of (11) is even $\Theta(\sqrt{n})$ with large probability over $X, Y$.

The same arguments yield that, at equilibrium of the chain, the width of the confidence interval (9) specified by (10) is of order $\epsilon\sqrt{n}$. Let for simplicity the prior be flat and the proposal symmetric, so that the condition in Step 8 of Figure 1 reads $\log u \in [b_-(\theta, \theta'), b_+(\theta, \theta')]$.

Since the middle of the confidence interval (9) is $\mathcal{O}_p(\sqrt{n})$, negative in a significant proportion of iterations, and its length is of order $\epsilon\sqrt{n}$ for "reasonable" datasets, the probability that once during the run $\log u$, minus an exponential variable, falls within the confidence interval (9), grows with $n$ unless $\epsilon = \mathcal{O}(n^{-1/2})$. Avoiding RPMH to fail thus requires $\epsilon = \mathcal{O}(n^{-1/2})$. In order to keep the probability in (8) small, this amounts to using $k = \Omega(n)$ random projections, thus violating our memory constraints.

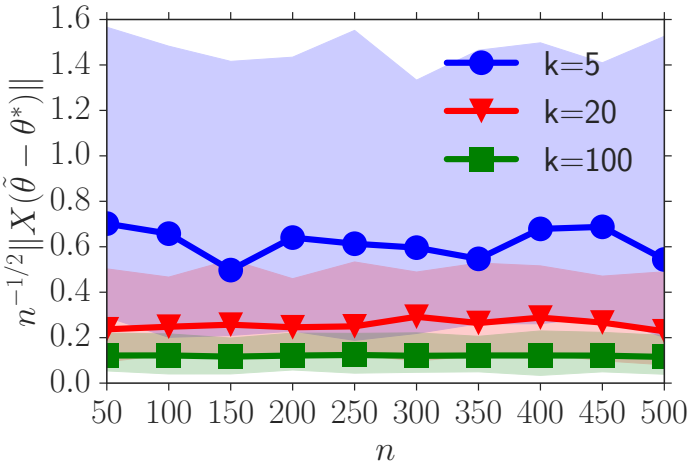


FIGURE 2. *Rescaled distance between $\theta^\star$ and $\tilde{\theta}$. 100 datasets were simulated from a linear Gaussian model for each pair $(k,n)$, with $X$ drawn from a spherical Gaussian. For each $k$, we show the median distance and the interquantile region between 10% and 90%.*

**Claim 1.** *When using RPMH with the naive confidence bound (10) on datasets $X, Y$ that come from a regression model with (non-degenerate) Gaussian noise, one needs to use $\epsilon = \mathcal{O}(n^{-1/2})$ and $k = \Omega(n)$ random projections in order to ensure that RPMH does not fail with high probability.*

**Using pivots.** The heuristics in this section are based on concentration of measure arguments, and are thus very similar to the one in (Bardenet *et al.* , 2014, Section 3.2.2), which says that subsampling-based MCMC needs $\Theta(n)$ points per iteration at equilibrium. In the case of subsampling, (Bardenet *et al.* , 2015, Section 7) show that one can go below $\Omega(n)$ with one pass over the data by remembering the MLE of the problem in "favourable" scenarios. The equivalent technique in our case would be to apply Proposition 3.1 to

$$\rho(\theta,\theta') = \langle X(\theta-\theta'), X(\theta+\theta'-2\theta^\star)\rangle.$$

The bounds in (10) are then replaced by

$$\begin{aligned} b_\pm(\theta,\theta') &= \langle AX(\theta-\theta'), AX(\theta+\theta'-2\theta^\star)\rangle \\ &+ \frac{\epsilon}{(1\mp\epsilon)^2}\|AX(\theta-\theta')\|\|AX(\theta+\theta'-2\theta^\star)\|. \end{aligned} \tag{13}$$

Since for $\vartheta = \theta, \theta'$, at equilibrium, $\|X(\vartheta-\theta^\star)\| = \mathcal{O}_p(1)$ by (12) and the choice of a proposal with inverse covariance $(AX)^T(AX)$, the same arguments as for the naive case yield that the width of the confidence interval (9) is now $\mathcal{O}_p(\epsilon)$. Thus $\epsilon = o(1)$ is enough to guarantee that $\log u$ only falls within the confidence interval (9) with arbitrarily small probability, and $k$ can be chosen as $o(n)$ while keeping the probability in (8) large.

**The $\theta^\star$ issue.** Unfortunately, having access to $\theta^\star$ means we can solve the normal equations. In particular we can compute products with $(X^TX)^{-1}$, and thus compute the exact log likelihood ratio (6) exactly. If this is the case, then there is no need for random projections in the first place. As stated in Section 1, we are more interested in cases where accurately computing $X^TX$ is not possible, say data is acquired in a turnstile manner and one only has $o(n)$ memory. In classical –as opposed to Bayesian– regression (Clarkson & Woodruff, 2009), the MLE $\tilde{\theta}$ of the projected problem is used as a proxy for $\theta^\star$.

Unfortunately, replacing $\theta^\star$ by $\tilde{\theta}$ in (13) requires to control $\|X\tilde{\theta} - \theta\|$. At equilibrium, $\theta$ will be close to $\theta^\star$ in the sense that $\|X(\theta - \theta^\star)\| = \mathcal{O}_p(1)$. However $\theta^\star$ will generally be far from $\tilde{\theta}$. In Figure 2, we empirically show that for $X, Y$ generated from the linear Gaussian model, $\|X(\theta^\star - \tilde{\theta})\|$ is $\Theta(\sqrt{n})$. This leads to the same scaling $\epsilon = o(n^{-1/2})$ as for the naive approach (10), which in turn leads to $k = \Omega(n)$ in Proposition 3.1 and the failure of the attempt to solve Bayesian linear regression in $o(n)$ memory.

### 4.2. A puzzling lower bound

To finish, we mention a lower bound on the memory budget that *any* randomized *one-pass* algorithm satisfies if it solves linear regression.

**Theorem 4.1.** *(Clarkson & Woodruff, 2009, Theorem 3.7) Let $\epsilon > 0$. Assume $n \geq d\log_{10}(nd)/(36\epsilon)$ and $d$ is sufficiently large. If a randomized 1-pass algorithm outputs, for any fixed $X, Y$ of size $n \times d$ and $n \times 1$, a value $\bar{\theta} \in \mathbb{R}^d$ such that*

$$\|X\bar{\theta} - Y\| \leq (1+\epsilon)\|X\theta^\star - Y\| \tag{14}$$

*with probability at least $7/9$, then this algorithm requires at least $\Omega(d^2\epsilon^{-1}\log(nd))$ bits of space.*

While Theorem 4.1 clearly applies to classical random projection regression, it is not immediately clear how to rigorously apply this lower bound to the RPMH algorithm in Figure 1. In particular, the proof of (Clarkson & Woodruff, 2009, Theorem 3.7) uses the fact the randomized algorithm solves (14) for every $X, Y$. If $Y$ is very close to the columnspace of $X$, then returning a uniformly drawn sample from the output of RPMH will typically not satisfy (14). But, for "reasonable" $X, Y$, say e.g. when the linear model is accurate and the distribution of the noise $\eta^\star$ is Gaussian, (14) will be satisfied with $\epsilon = n^{-1}$, again leading to $k = \Omega(n)$.

### 4.3. On previous work

Geppert *et al.* (2015) study Bayesian regression with random projections and show that with large probability on the JL matrix $A$, the posterior of the projected regression

$$\tilde{\pi}(\theta) \propto \exp(-\|AX\theta - AY\|^2)p(\theta)$$

is within $\mathcal{O}(\epsilon^2)$ of the posterior $\pi$ in 2-Wasserstein distance. This is an interesting result, but being close in Wasserstein distance is not sufficient to build credible intervals on $\theta$ –a common task in Bayesian inference– using less than $\Omega(n)$ random projections.

Indeed, the leading term in (Geppert *et al.* , 2015, Lemma 8) comes from a bound

$$\|\theta^\star - \tilde{\theta}\|^2 = \mathcal{O}(\epsilon^2), \tag{15}$$

on the squared bias in (Geppert *et al.* , 2015, Lemma 6). Since both $\pi$ and $\tilde{\pi}$ put most of their masses on regions of diameter of order $n^{-1/2}$, we would need $\epsilon$ to be of order $n^{-1/2}$ in order for (15) to guarantee that credible intervals built with $\pi$ and $\tilde{\pi}$ do overlap. Note that requiring that the credible intervals overlap is a minimum if $\tilde{\pi}$ is to be used to make inference on $\theta$. Again, having $\epsilon$ of order $n^{-1/2}$ requires $k = \Omega(n)$ in JL-type results such as Proposition 3.1, and the $o(n)$ memory constraint is not satisfied.

## 5. Conclusion

Random projections are undeniably useful for classical linear regression, even under constrained data acquisition. It is natural to wonder whether random projections can help approximate the log acceptance ratio in MCMC for Bayesian linear regression. Uniform JL-type results guarantee that random projection confidence bounds can hold tight across MCMC iterations, a feature not accessible to uniform subsampling. However, we have explained why making MCMC acceptance decisions based on such bounds requires a higher degree of accuracy than what $o(n)$ random projections can provide. Control variates such as the MLE of the projected problem cannot fundamentally help.

Future work should investigate communication complexity results in the line of (Clarkson & Woodruff, 2009), to derive a rigorous lower bound on memory complexity of one-pass MCMC for linear regression. Another interesting avenue is to look for better control variates than the MLE of the projected problem, possibly relaxing the constraint of doing being one-pass, such as in Zhang *et al.* (2013), as long as memory remains controlled. Whether there is a fundamental trade-off or enough room for a win-win method is still an open question.

*Acknowledgements*
This work is supported by the French national center for scientific research (CNRS) and INS2I under PEPS grant PROMO, by Inria, and the LRI and CRIStaL laboratories.

## References

Achlioptas, Dimitris. 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, **66**(4), 671–687.

Ailon, Nir, & Liberty, Edo. 2009. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, **42**(4), 615–630.

Alon, Noga, Goldreich, Oded, Håstad, Johan, & Peralta, René. 1992. Simple Constructions of Almost k-wise Independent Random Variables. *Random Structures & Algorithms*, **3**(3), 289–304.

Alon, Noga, Matias, Yossi, & Szegedy, Mario. 1996. The space complexity of approximating the frequency moments. *Pages 20–29 of: Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. ACM.

Alon, Noga, Andoni, Alexandr, Kaufman, Tali, Matulef, Kevin, Rubinfeld, Ronitt, & Xie, Ning. 2007. Testing k-wise and almost k-wise independence. *Pages 496–505 of: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM.

Baraniuk, Richard G. 2007. Compressive sensing. *IEEE signal processing magazine*, **24**(4).

Bardenet, R., Doucet, A., & Holmes, C. 2014. Towards scaling up MCMC: an adaptive subsampling approach. *In: Proceedings of the International Conference on Machine Learning (ICML)*. http://jmlr.org/proceedings/papers/v32/bardenet14-supp.pdf.

Bardenet, R., Doucet, A., & Holmes, C. 2015. On Markov chain Monte Carlo methods for tall data. *Preprint, available as http://arxiv.org/abs/1505.02827*.

Boucheron, S., Lugosi, G., & Massart, P. 2013. *Concentration inequalities: a nonparametric theory of independence*. Oxford University Press.

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (eds). 2011. *Handbook of Markov Chain Monte Carlo*. CRC Press.

Candes, Emmanuel J, & Tao, Terence. 2006. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, **52**(12), 5406–5425.

Clarkson, K. L., & Woodruff, D. P. 2009. Numerical Linear Algebra in the Streaming Model. *In: Proceedings of the ACM Symposium on Theory of Computing (STOC)*.

Dasgupta, Sanjoy, & Gupta, Anupam. 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random structures and algorithms*, **22**(1), 60–65.

Donoho, David L. 2006. Compressed sensing. *Information Theory, IEEE Transactions on*, **52**(4), 1289–1306.

Duarte, Marco F, Davenport, Mark, Wakin, Michael B, Laska, Jason N, Takhar, Dharmpal, Kelly, Kevin F, Baraniuk, Richard G, *et al.* . 2007. Multiscale random projections for compressive classification. *Pages VI–161 of: Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 6. IEEE.

Geppert, L. N., Ickstadt, K., Munteanu, A., Quedenfeld, J., & Sohler, C. 2015. Random projections for Bayesian regression. *Preprint, available as http://arxiv.org/abs/1504.06122*.

Gilbert, Anna C, Strauss, Martin J, Tropp, Joel A, & Vershynin, Roman. 2007. One sketch for all: fast algorithms for compressed sensing. *Pages 237–246 of: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM.

Hamilton, William L, Fard, Mahdi M, & Pineau, Joelle. 2013. Modelling sparse dynamical systems with compressed predictive state representations. *Pages 178–186 of: Proceedings of the 30th International Conference on Machine Learning (ICML-13)*.

Huang, Z., & Gelman, A. 2005. *Sampling for Bayesian computation with large datasets*. Tech. rept. Department of Statistics, Columbia University.

Johnson, W. B., & Lindenstrauss, J. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, **26**, 189–206.

Korattikara, A., Chen, Y., & Welling, M. 2014. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. *In: Proceedings of the International Conference on Machine Learning (ICML)*.

MacLaurin, D., & Adams, R. P. 2014. Firefly Monte Carlo: Exact MCMC with Subsets of Data. *In: Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*.

Maillard, Odalric-Ambrym, & Munos, Rémi. 2012. Linear Regression with Random Projections. *Journal of Machine Learning Research*, **13**, 2735–2772.

Minsker, S., Srivastava, S., Lin, L., & Dunson, D. 2014. Scalable and Robust Bayesian Inference via the Median Posterior. *In: Proceedings of The International Conference on Machine Learning (ICML)*.

Muthukrishnan, S. 2005. *Data streams: algorithms and applications*. Foundations and Trends in Theoretical Computer Science.

Neiswanger, W., Wang, C., & Xing, E. 2014. Asymptotically exact, embarassingly parallel MCMC. *In: Proceedings of the conference on Uncertainty in Artificial INtelligence (UAI)*.

Quiroz, M, Villani, M., & Kohn, R. 2014. Speeding Up MCMC by Efficient Data Subsampling. *Preprint, available as http://arxiv.org/abs/1404.4178*.

Quiroz, M, Villani, M., & Kohn, R. 2015. Scalable MCMC for Large Data Problems using Data Subsampling and the Difference Estimator Speeding Up MCMC by Efficient Data Subsampling. *Preprint, available as http://arxiv.org/abs/1507.02971*.

Robert, C. P., & Casella, G. 2004. *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Roberts, G. O., & Rosenthal, J. S. 2001. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **16**, 351–367.

Sakai, Tomoya, & Imiya, Atsushi. 2009. Fast spectral clustering with random projection and sampling. *Pages 372–384 of: Machine Learning and Data Mining in Pattern Recognition*. Springer.

Sarlós, T. 2006. Improved Approximation Algorithms for Large Matrices via Random Projections.

*In: Proceedings of the IEEE Symposium on the foundations of Computer Science (FOCS).*

Scott, S. L., Blocker, A. W., & V., Bonassi F. 2013. Bayes and Big Data: The Consensus Monte Carlo Algorithm. *In: Proceedings of the Bayes 250 conference.*

Srivastava, S., Cevher, V., Tran-Dinh, Q., & Dunson, D. B. 2014. WASP: scalable Bayes via barycenters of subset posteriors. *Preprint.*

Wang, Wei, Garofalakis, Minos, & Ramchandran, Kannan. 2007. Distributed sparse random projections for refinable approximation. *Pages 331–339 of: Proceedings of the 6th international conference on Information processing in sensor networks.* ACM.

Welling, M., & Teh, Y. W. 2011. Bayesian Learning via Stochastic Gradient Langevin Dynamics. *In: Proceedings of the International Conference on Machine Learning (ICML).*

Zhang, Lijun, Mahdavi, Mehrdad, Jin, Rong, Yang, Tianbao, & Zhu, Shenghuo. 2013. Recovering the Optimal Solution by Dual Random Projection. *Pages 135–157 of: Conference on Learning Theory.*